Appendix

In the appendix, we mainly provide quantitative and qualitative results of our method and the state-of-the-art camera-based SSC method MonoScene [1] on the hidden test set of SemanticKITTI [2]. Since we do not have access to the ground truth of the test set, we can only report the performances within the full range ($51.2\times51.2\times6.4\text{m}^3$).

A. Quantitative Comparison

Scene completion. As shown in Table I, VoxFormer outperforms MonoScene with a large gap in terms of geometric completion. VoxFormer-S without using historical observations improves MonoScene on IoU with a relative gain of 25.73%. Note that in autonomous driving, geometry occupancy is critical for obstacle avoidance since a false negative could result in severe accidents. Therefore, our method is more desirable than MonoScene in safety-critical camera-based autonomous driving applications.

Semantic scene completion. As shown in Table I, VoxFormer also demonstrates a better semantic scene understanding. VoxFormer-S and VoxFormer-T both demonstrate better mIoU than MonoScene. VoxFormer-T / VoxFormer-S have a relative improvement of 21.03% / 10.11% compared with the cutting-edged MonoScene. Note that the values of IoU and mIoU are intertwined, and some methods can naively increase the value of mIoU by sacrificing IoU. In contrast, our method shows superior performance in terms of both geometry and semantics.

Short-range performances. Although short-range evaluations are not available on the hidden test set, we expect to see a similar trend (we perform much better in safety-critical short-range areas than MonoScene). The reason is that the scores of mIoU and IoU on the test set are comparable to that on the validation set inside the $51.2\times51.2\times6.4\text{m}^3$ volume. For example, VoxFormer-S achieves an mIoU of 12.35 on the validation set and 12.20 on the test set.

B. Qualitative Comparison

More visualizations are shown in Fig. I. We can see that our method performs much better than MonoScene in the short-range areas. There are some missing objects for MonoScene at close range, as shown in the first and last row of Fig. I. Meanwhile, the long-range performance of our method can be further improved, *e.g.*, the trunks in the long-range areas are not completed in the fourth row of Fig. I.

References

[1] Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3991–4001, 2022. 1

[2] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9297–9307, 2019. 1

Method	IoU	car (3.92%)	bicycle (0.03%)	motorcycle (0.03%)	truck (0.16%)	other-veh.(0.20%)	person (0.07%)	bicyclist (0.07%)	motorcyclist (0.05%)	road (15.30%)	parking (1.12%)	sidewalk (11.13%)	other-grnd(0.56%)	building (14.10%)	fence (3.90%)	vegetation (39.3%)	trunk (0.51%)	terrain (9.17%)	pole (0.29%)	traf.-sign (0.08%)	mIoU
MonoScene	34.16	18.80	0.50	0.70	3.30	**4.40**	1.00	1.40	**0.40**	**54.70**	24.80	**27.10**	5.70	14.40	11.10	14.90	2.40	19.50	3.30	2.10	11.08
VoxFormer-S (Ours)	42.95	20.80	1.00	0.70	3.50	3.70	1.40	**2.60**	0.20	53.90	21.10	25.30	5.60	19.80	11.10	22.40	7.50	21.30	5.10	4.90	12.20
VoxFormer-T (Ours)	**43.21**	**21.70**	**1.90**	**1.60**	**3.60**	4.10	**1.60**	1.10	0.00	54.10	**25.10**	26.90	**7.30**	**23.50**	**13.10**	**24.40**	**8.10**	**24.20**	**6.60**	**5.70**	**13.41**

Table I. Quantitative results of VoxFormer and the state-of-the-art MonoScene on the hidden test set of SemanticKITTI.

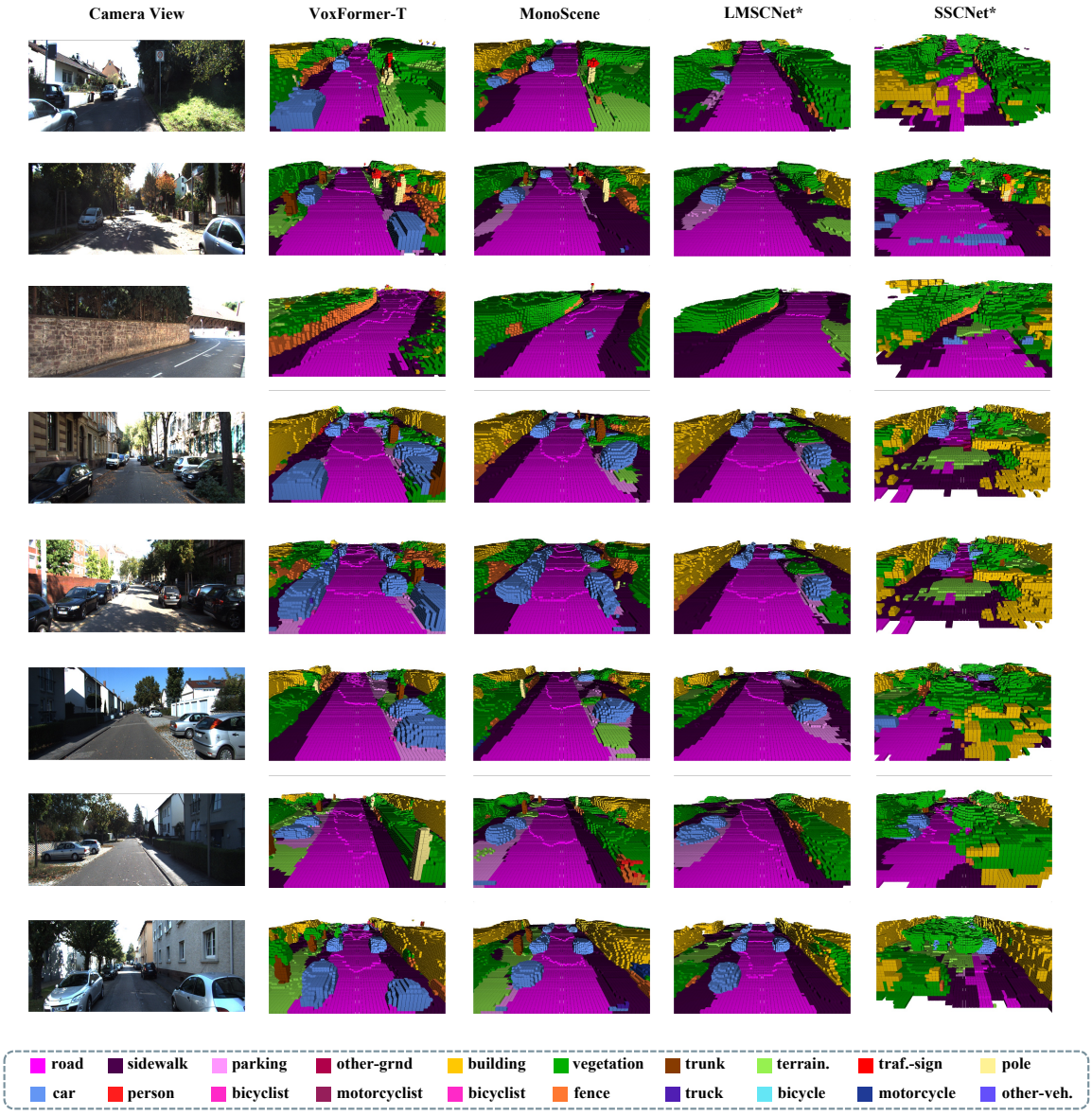

Figure I. **Qualitative results of our method and others on the hidden test set.** VoxFormer better captures the scene layout in large-scale self-driving scenarios. Meanwhile, VoxFormer shows satisfactory performances in completing small objects such as trunks and poles.